DOCUMENT RESUME

ED 220 963 EA 014 948

AUTHOR Troob, Charles; Ryzewic, Susan Remmer
TITLE Accuracy of Data in a Computerized Student Record System for a Large District.
PUB DATE 20 Mar 82
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Oriented Programs; Confidentiality; *Databases; Data Processing; Education Service Centers; Elementary Secondary Education; Information Needs; Management Information Systems; Public Schools; *Quality Control; Recordkeeping; *Standards; *Student Records
IDENTIFIERS *Integrated Student Data System; *New York City Board of Education

ABSTRACT

The New York City Public School System has mandated an Integrated Student Data System (ISDS). This integrated system will include data already maintained in independent computerized files, as well as data that have not been computerized in any systematic way. This paper explores general issues about data accuracy for a system that covers all the students in the school system and that will include information required for a very wide range of functions. The paper develops three major themes. The first is that the need for data accuracy depends upon the use for which the data are intended. A second theme is that data system participation must be encouraged, for instance by maximizing the "fit" between the data system and the day-to-day operation of the school. The final theme is that the ISDS must ensure timely submission by schools of accurate data, that performance must be monitored, and that it must be possible to provide feedback. For each subject, discussion of issues and complications yield operating principles and guides for action. (Author/MLF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED220963

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Charles Troob

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ACCURACY OF DATA IN A COMPUTERIZED STUDENT RECORD SYSTEM FOR A LARGE DISTRICT

Charles Troob
Office of Student Information Services
New York City Public Schools

Susan Remmer Ryzewic
Office of Academic Affairs
City University of New York

Presented at the American Educational Research Association
New York
March 20, 1982

EA 014 948

I. INTRODUCTION

This study of accuracy of student data has its source in a number of activities of the New York City Public Schools. For several years, the school system has been developing a student locator file (now called the Biofile) as the first stage in the implementation of an Integrated Student Data System (ISDS). This integrated system will include data already maintained in independent computerized files, as well as data which has not been computerized in any systematic way. The system is intended to encompass a number of current and planned efforts in the processing of student data, ensuring that redundant activities are avoided, and that services can be provided which make the maximum use of whatever efforts are devoted to data collection and processing.

The Biofile is now operational. Schools submit monthly reports of all students admitted and discharged. These are submitted on ledger sheets from all elementary and middle schools, and many high schools. About half of our high schools have contracted with the City University for the services of a computerized scheduling, grade reporting and daily attendance package. Data from these schools are submitted by the processing center in machine-readable form. The Division of High Schools is operating a pilot project to provide similar services in-house with purchased software, and the two schools in this pilot are also submitting machine-readable data.

The Biofile now has the elements of a regular monthly processing cycle: input documents, error reports, and periodic listings of the file for use in updating biographical and class information. A historical file is maintained, which includes all school location changes back to September, 1981. The historical

data are used in the editing of transactions, as the system cannot take for granted that moves will be reported in the order in which they take place.

The Biofile includes student name, sex, date of birth, ID number, home address, school, grade, class, and some codes. It has a number of direct applications, and many more when merged with files coded by ID number which contain additional data. The plan for data system development calls for a wide range of additional information to be added to the basic Biofile data, and to make the entire data system more accessible to authorized users throughout the public school system.

In 1980-81, the general requirements and strategy of the Integrated Student Data System were set out in a document entitled Plan for an Integrated Student Data System (New York City Public Schools). That report is both a policy paper and a specifications document. It reflects the belief of many people in the school system that a data system is both a technical apparatus and an important institutional change. To be successful, the designers of the data system will have to pose the institutional problems in the correct way, and then to find technical solutions to them.

One of the major institutional problems is considerable skepticism concerning the feasibility of the ISDS. A number of past computerization efforts have been at best partially successful, and school personnel are naturally suspicious of any new effort which promises a great deal of clerical work and no very certain payoff. Parents are justifiably concerned about the creation of data records on their children. Parental concern was an important reason for the creation of a Chancellor's Committee on Confidentiality, which has the right to mandate procedures governing access to student records, as well as the right to veto

the creation of data files which the Committee judges to be inappropriate in content or planned use, or inadequately safeguarded.

During the planning effort, and in discussions with the Confidentiality Committee, one issue that was repeatedly raised was the difficulty of ensuring the accuracy of data in any data system which serves as varied a clientele as the schools of New York City. New York City has nearly 1000 public schools, and the elementary and middle schools are governed through a system of Community School Districts. Although the system was once rigidly centralized, and most key procedures are centrally mandated, there is considerable local variation in how data are maintained and reported. In addition, the system for delivering special education services is changing and expanding rapidly.

In such an environment, and with expanding concerns over privacy and civil liberties in the computer age, the public school system must demonstrate that it can keep student data both safe and accurate.

All data system planning must deal with the issue of accuracy. Systems designers are expected to pilot their input documents, to develop extensive controls over paper flow, and to build edit routines into their data entry programs. The Integrated Student Data System requires all these activities. The Office of Student Information Services, responsible for the ISDS, has considerable experience with these tasks, and gains more with every month.

Without in any way denigrating the importance of these absolutely essential data management functions, we would like here to explore the more general issues about data accuracy which derive from the mandate of the ISDS: that it be a data system about students; that it include all students in the New York City Public

Schools; and that it include the information required to perform a very wide range of functions.

The paper is an outgrowth of questions raised during the planning phase, and also is an interpretation of the experience of the school system with a number of data processing activities: the Biofile itself, computerized test scoring, a Competency Test history file, evaluation reporting, the Data Bank of special education students, the transportation data base, the extensive high school data processing services, and activities in individual Community School Districts. The authors also investigated other operational student data systems.

The paper develops three major themes. The first is that the need for data accuracy depends upon the use for which the data are intended. Data collected with a particular set of uses in mind may not be satisfactory for other uses. This has a range of implications for a system whose objective is to put data to many uses. In particular, it affects the rate at which new data elements should be phased in and the particular applications of the data.

A second theme is that data system participation must be encouraged by maximizing the "fit" between the data system and the day-to-day operation of the school. As much as possible, data submissions should be part of regular school operations, and the correction and maintenance of data files should be a by-product of school use of the data system or its products. While this is a major objective, it is important to recognize that the ISDS will have to contain data that is currently not maintained at all, or else is maintained with considerable error in many schools. It will be necessary for schools to develop or modify operating procedures if they are to maintain such data to the level of accuracy suitable for any data system use.

The final theme is that the ISDS must be prepared to ensure compliance, by which we mean timely submission of accurate data. As in any administrative system, performance must be monitored, and it must be possible to provide feedback. When feedback does not suffice, stronger methods of enforcing compliance must be available. This is an unpleasant issue to face when discussing something which is billed as a "service," but it is an inherent requirement in view of the system mandate. Many of the proposed uses of the ISDS will not be possible unless all schools fulfill their responsibilities.

In short, this paper is about three subjects: how good the data have to be; how to encourage schools to submit and maintain data to that standard; and what to do if they don't. For each subject, we have tried to take an almost unmanageable number of issues and complications and turn them into a few operating principles and guides for action. If we have been successful, this paper, which is in no sense an official planning document, will have some impact on the further planning and development of the ISDS.

II. ACCURACY OF DATA FOR THE ISDS--AN OVERVIEW

Categories of data and methods of collection

The first categories of data to be maintained in the Integrated Student Data System are student characteristics, location in the school system, test results, attendance, and program participation. Much but not all of this data is currently maintained manually in the schools.

The method of data submission and entry will evolve as the system develops. Initially, all entry will be done centrally, except for those schools currently served by data systems which can be linked with the ISDS. As soon as possible, the system will have terminals in district offices and in schools. These will first be used to inquire on data, and then for data entry. Eventually, most data will be entered directly onto the system from schools and other field locations.

For the next few years, data will be submitted on forms, either to a single processing center or to regional centers. These forms will include key-entry forms and optical scan sheets. Figure I is a table from the Plan which describes the most probable method for collecting different kinds of data.

In general, school secretaries will have the bulk of the work involved in interfacing with the data system, as they are responsible for maintaining information about school location, and for collecting data on new entrants. The system will also require information, such as testing and program data, which can only be supplied by pedagogic staff. Thus a number of different school personnel will be providing information to the data system, and making use of system products.

FIGURE I

Need for Forms for Data Submission to the ISDS - Phase I

Purpose	Type of Form
1. Admission of new students	1. Individual form or group list
2. Correction of information (e.g. personal information, test history)	2. Individual form -- possibly pre-printed or Pre-printed roster with correction field
3. New data field for all students (e.g. immunization, program participation, cumulative attendance)	3. Scannable roster or form if possible; otherwise pre-printed roster for key-entry
4. School changes - individual	4. Individual form or group list.
5. School changes - articulation	5. Enter new school on pre-printed list of students in terminal grades (perhaps scannable); or scan student forms along with destination school header sheet
6. Class/grade changes -- mass change	6. Not yet decided. Machine readable system would be preferable. If not possible, enter on pre-printed roster
7. Class/grade changes -- individual	7. Individual form or group list (see 4)
8. New data field -- some students potential holdovers, no shows, test results from tests not scored by OSIS.	8. Pre-printed roster, if possible, or list
9. Eliminate duplicate ID's	9a. Individual form or list - if school discovers duplicate b. Pre-printed form -- if system discovers duplicate
10. Collect ID numbers for students whom system cannot identify (e.g. after tests have been given)	10. Pre-printed roster

Note: Tests taken by all or virtually all students can be used as data data collection instruments (e.g. to update official class).

Sources of error

A strategy for data accuracy must take into acccunt a very wide range of reasons for error. Different kinds of error imply different methods of avoidance, discovery, and correction, and also have different implications for the basic integrity of the record system.

While many different typologies of errors are possible, the following taxonomy has been particularly helpful in clarifying issues concerning data for the ISDS.

A key reason for data system error is the school's failure to maintain correct information in the first place. This failure can arise from many quite different sources. Schools may consider the information of low priority, and fail to collect it although it would be possible to do so. (A particularly annoying example of this is the failure of the school to obtain the correct ID number, assigning the student a new one instead.) In other cases, schools may perform the information-gathering activity with such sloppy procedures that the data are unusable for or inaccessible to the record system. Schools with great amounts of mobility may not be able to collect information on some incoming students, given the limited resources of any school for non-pedagogic activity. Finally, the student and family may have deliberately submitted inaccurate personal information, and the school may not have performed appropriate verification techniques.

Most cases of deliberate misreporting by schools fit into this category, as the school also fails to maintain correct information internally: answer sheets not submitted for students who were supposed to have taken tests; incorrect addresses for out-of-zone students; students reported as program participants who attended no

more than once; students inappropriately kept on register. (In some cases, of course, the school does have the correct information somewhere on file.)

What distinguishes this entire category of error is that the failure to maintain correct information is not an error of the computerized system per se, but of the schools' record-keeping procedures in general. Thus the focus of attention must be on schools' record-keeping behavior.

Because the "errors" are due to circumstances largely out of control of the data system, this type of error raises the most subtle and complex issues in data system design. A computer system can contribute to reducing this kind of error: by prompting the schools for missing data, by computerized checks on consistency, by maintaining data on mobile students, and by facilitating audits. In fact, the computer's main contribution to record-keeping lies in reducing this type of error. On the other hand, a computer is no panacea, and data maintained badly in a good manual system is not likely to be well handled in a computerized one.

A second major reason for error is the failure to report information which the school possesses. In its most blatant form, this is due to deliberate and open non-compliance with data system requirements.

Except in this case, non-reporting is usually due to problems in system design, such as the lack of exact congruence between school administrative procedures and information system data collection procedures. For example, unless pre-printed data collection forms or rosters are arrayed in official class order, schools will find it burdensome or impossible to determine which students are missing and must be reported on a supplementary sheet. In effect, this requires the data system to collect and maintain

information on official class, although for most purposes official class information is not necessary on a centralized file, and the information must be frequently updated.

An analogous problem is the failure to report data out of the normal reporting cycle: a test score raised after appeal, or a make-up exam. Failure to resubmit information which was originally rejected (perhaps because an ID number was wrongly entered) is another important form of non-submission.

Terminal entry in schools will be helpful in avoiding this category of error, as schools with terminals will find it relatively easy to maintain accurate class lists on file and to enter small amounts of supplementary data. However, the placement of terminals in schools will not make these things happen automatically. Schools must still have a strong incentive to keep their data accurate and up to date.

A third category of error is simply the incorrect submission of data. This takes many forms: transcription error; illegibility; the misuse of codes; and the more general failure to follow instructions. Incorrect submission is often a more subtle form of non-compliance, which reflects the school's willingness to follow orders but not to do the job well. However, what is often interpreted as non-compliance may be due to misjudgment in design or poor implementation. Inadequate notice, inadequate training, and poorly designed forms and procedures can all be the indirect causes of poor data submission. Conversely, well-designed procedures are the best insurance against poorly submitted data.

A fourth category is error in processing. This category includes loss of data in transmission from school to processing center, keypunch error, or errors in processing programs. This

kind of error is the one that most data processing people are good at avoiding, or at least in telling you how to avoid. In other words, go by the book in testing programs, piloting internal procedures, logging, and data editing.

The final category in this chronological taxonomy is failure to discover and correct error after it has taken place. While it is possible to mandate scrutiny of data from time to time, the best insurance against this type of error is for the data to be used regularly by people who are able to evaluate its validity and to change it if erroneous. Thus accuracy depends heavily on the use of data by school personnel. However, as the next section discusses, it is important not to promote certain uses for which the data may not be sufficiently accurate.

III. THE NEED FOR ACCURATE STUDENT DATA IN DIFFERENT APPLICATIONS

A moment's reflection will suggest that the level of accuracy required of a data series will depend on how the data will be used. This obvious point is rarely acknowledged in planning or discussing a data system, perhaps because administrators and planners do not wish to condone the inevitable error which will remain after reasonable precautions have been taken. Yet administrators of data activities readily acknowledge the fact of error, often asserting that the error rate, however deplorable, still remains within an acceptable range.

This section attempts to establish the validity of two principles: (1) that a data series can be quite acceptable for some important uses, while being seriously deficient for other possible uses; (2) that it will take several years before the ISDS can keep any data series to the highest desirable level of accuracy. We present a simple categorization of data uses which is consistent with the first point, and an argument for the second point. The implication of these two principles is that the survival of a system like the ISDS depends on a clear acceptance of what it can be expected to attain in its first years, as well as steady progress toward higher standards of accuracy.

Once we put the question, "How accurate must the data be?" many possible factors come to mind. Are the data likely to be of permanent legal significance? What decisions about each student could the data possibly influence, and how important are these decisions? It a more accurate version of the information likely to be maintained? Is anyone likely to discover an error before harm is done? Will the data be used to generate or justify funding, and must we guard against the possibility of future disallowances? Will the data be used to generate important statistical series?

Will the data be used to determine whether programmatic responsibilities are being carried out?

It might be quite interesting to deal with the full complexity of these and similar questions. At the risk of oversimplification, we believe that it is very instructive to consider three broad categories of data use, which have distinctive needs for data accuracy. It should be noted that all three categories require data files at least as complete and accurate as those that researchers are accustomed to working with. We present the categories in order of the general level of accuracy which they require, from lowest to highest.

Accuracy Level 3: For statistical analysis, production of aggregates, quantitative program evaluation, charting of trends, assessment of school system problems and needs.

These uses of data, primarily statistical and analytic, do not require the absence of error in individual records, so long as the data are approximately correct. For some analytic uses, error rates of 5-10% are quite acceptable; for some statistical activities, the actual rate of error is of less importance than whether or not the errors are random and unbiased, or whether the biases can be allowed for or can be assumed constant over time. Program evaluators rarely have the luxury of data which are 95% complete and accurate.

For most analytic work, the reason to be concerned with errors of this magnitude is not that they make it impossible to come to valid conclusions, but that if the public were to know how much is incorrect the findings would lose much credibility. It is noteworthy that missing data seems more publicly acceptable than incorrect data, even though incorrect data is more likely to be

caused by random transcription errors, while missing data is more likely to cause bias in the findings.

The hallmark of these data uses is that the student is a "case" rather than an individual. The only reason to identify the student is so that complete records of each case can be obtained, and so that errors can be tracked down and files cleaned. The concern that erroneous information may lead to a wrong decision about an individual does not apply here. It is not uncommon for researchers trained in this way of looking at data to be quite bewildered when others reject "approximately correct" data.

Accuracy Level 2: For maintaining records of school operations, such as aggregate registers and attendance, program enrollees, services provided.

These uses of data are both statistical and individual. While these records are rarely the only source of important information about a student, they are used to judge whether or not schools are properly carrying out their functions, and whether the school system is justified in filing financial claims. The data must be credible as a record of day-to-day circumstances. Although for most uses the data are aggregated, the individual records may be examined by a supervisor, an auditor, or a program monitor. These individuals do not expect to find perfect data--audit disallowances of a few percentage points are a way of life--but a considerable degree of error is likely to invalidate the entire data series.

As a major concern of a monitor is face validity, it is important for data put to these uses to be free of the random errors which lead to glaring inconsistencies. It is also important that the data agree with other information. Because computer editing can automatically get rid of inconsistencies, auditors are likely to want to see "hard copy" data as well. Thus these uses

may limit the degree to which the data system can replace manual record-keeping.

The likelihood of audit is usually a key determinant of the care with which these kinds of records are kept. This is often hard to foresee, as audit agencies are not likely to advertise their plans. The possibility of external auditing generates a need for internal auditing, just to ensure that the data are adequate should the external audit ever happen.

Accuracy Level 1: For making determinations about individual students.

In general, individual record systems are perceived to exist primarily to support decisions about individuals, without reference to other sources of information. Of course, many (if not most) school decisions involve complex factors, and are not data-determined. But many decisions do rely heavily or exclusively on the value of certain data elements. A student's age determines eligibility for school. Address determines school zone or eligibility for open enrollment programs, as well as Title I eligibility. (Address also determines Impact Aid eligibility, but this is a Level 2 concern.) Test scores now determine eligibility for promotion, and the New York State Regents Competency requirements for high school graduation are comprised of test results. Test scores and attendance (as well as course grades) are extensively used in high school admissions and in placement into a wide range of programs. A previous year's participation in a program may be a criterion of present eligibility. The absence of adequate documentation of immunization is (periodically) sufficient reason to bar students from school.

How accurate must data be to support Level 1 uses in a way consistent with anyone's definition of student rights and due process? If data are to be used in such ways, three key considerations affect the need for accuracy: Will all such uses be performed by individuals who have personal knowledge of the student? Will the fact that a decision has been made be explicitly acknowledged, and the data cited as the reason for the decision? Does the school or the family also maintain other, more accurate records which can be appealed to?

If the answer to two or three of these questions is "yes," the probability is high that mistakes will be discovered and rectified before any damage is done, so that accuracy is less critical. Such in-school applications of data as class assignment or issuance of transportation passes often meet this criterion. These in-school uses of data are not readily distinguishable from those operational activities which we have characterized in the discussion of Level 2, and perhaps no greater accuracy is required for these "Level 1C" applications. (Keep in mind that an data set worthy of external audit is a very clean data set indeed.)

In many schools, however, particularly large ones, data is likely to be used uncritically in scheduling students for special programs or tests, or in competitive admissions procedures. If these activities are performed by people unfamiliar with the student, data errors may cause mistakes whcih never come to light. If a student changes schools, almost any errors made by the previous school are hard to resolve. Thus any data which may affect decisions in a school other than the one submitting the data must be kept extremely accurately, at "Level 1B."

Finally, some student records are among an individual's most precious possessions. Major mistakes in a student's permanent file (or its disappearance) can be harmful, and irremediable later in life. School records are used to certify age, identity, and residence history in a wide variety of legal settings, including eligibility for Social Security at age 65. The adoption of the

High School Equivalency program has perhaps lessened the long-term significance of the transcript, but it is still very important to applicants to jobs and higher education. The maximum effort possible is required to avoid errors in the permanent record of some categories of information about students. Let us call the accuracy level suitable to such crucial information Level 1A.

Discussion:

Usually, administrators set out to assemble "correct" data. The distinction made here among levels of accuracy exists only at the back of their minds, if at all. Yet their actions usually reflect some judgments about how much accuracy is really necessary. A data collection activity may be conceived in such a way as to guarantee some missing data; or they may fail to devote the resources to piloting, training, monitoring, and editing, which would be required to achieve Level 1 data.

While sometimes this result can be ascribed to inexperience or negligence, it often is due to a good sense of what is really necessary for the task at hand, the one that motivated their activities. Any single-purpose data collection includes a number of data elements, some of which must be much more accurate than others. Often the limitations of the data are not discovered until someone else tries to put them to a quite different use.

Data collection is an "economic" activity like any other, with results related to the level of input. What is not sufficiently recognized is that high quality data is much more costly to produce than low quality data. Survey researchers are familiar with the notion that the first 90% of the "universe" may be much easier to reach than the remaining 10%, and particularly the last 1%. Quality controllers are well aware that it costs increasingly more to get work done to a more and more exact standard. Psychologists know that there is a learning curve, and

that anything done the first time is likely to be done less well than after it has been done repeatedly.

One reason why this is so often overlooked is that school record-keeping is such a traditional activity. We tend to take for granted the amount of sensible technology built into the cumulative record and the rollbook, and also into the conventional procedures which govern the taking of attendance and the recording of grades. In schools which conform to these procedures, basic student records are kept to an extremely high standard, and attitudes generated by these record systems spill over into newer activities. Many of our administrators insist that any data set involving students be kept to very high standards. These are the people who take clerical work home, fanatically scrutinize the efforts of those they know to be less dedicated, and double-check every printout they receive.

That such people can produce good data under almost any circumstance should not blind us to the difficulty of getting uniformly good data. These people are clearly exemplary, and not every school has even one. We cannot take for granted that our data will be supplied by people with enormous concern for its accuracy.

The foregoing discussion suggests that the Integrated Student Data System will only in time become as accurate and complete as our standard for permanent records. For this reason, if no other, the need for hard copy will exist for years after the implementation of computerized files. Thus we should not promise the actual elimination of manual records in the near future. A computerized data system will for a number of years be a support system for a record system which remains manual.

As the next section explains, a computer system supports a manual system in many ways. The applications discussed in that section require very good data, in the sense that researchers understand this term, but none of them requires that the ISDS

become the system of record for any data element at all, except possibly the New York City ID number. By "system of record," we mean the system that overrules any other information sources maintained in schools. For example, the data system can be used to print out registers for use in auditing enrollment, and to resolve issues of who was where on a given day, without having to maintain that in cases of discrepancy the data system is always right.

In general it will be impossible to keep the promise that a data system will make it unnecessary for schools to keep a paper record (if only a printout or label) of an important category of student information. Such promises should not be made for some time to come. In addition, it is probably unadvisable to ask schools to rely on the data system without further verification when making particularly important decisions about students. Applications that require very high levels of accuracy should be the last to be implemented.

Another implication of this line of analysis is that there may be multiple Level 3 uses for student data which do not justify the effort involved in getting the data better than approximately correct. What are we to do about data than are only used for Level 3 uses? There is a risk that, after collection, someone will (inappropriately) take data collected for statistical purposes and put it to administrative uses, or even for decisions about individuals. This is what parents rightly fear.

There are two possibilities here. One could take the hard line, and argue that "approximately good" isn't good enough as applied to student information. This would suggest that data never be maintained unless they are found to be of Level 2 quality or better. A disadvantage of this approach is that it encourages administrators to exaggerate the quality of their data in order to continue using them for purposes for which they are good enough.

A less drastic strategy would be to identify clearly those data elements which are known to be only "approximately correct,"

in such a way as to prevent or at least forestall their use for program monitoring or the answering of questions about individuals. They might be suppressed on on-line display, and printed on reports along with warnings that they require further verification.

IV. MATCHING THE SCHOOL SYSTEM AND DATA SYSTEM TO EACH OTHER

Schools use data for a variety of purposes. (1) They fulfill legal responsibilities in the maintenance of student permanent records. (2) They assemble data to manage the school on a day-to-day basis, in such areas as attendance, scheduling, promotion and grading, the provision of supportive educational services, and the provision of food, health, transportation and other services. (3) They maintain auditable records on all their activities. (4) They submit reports which summarize data collected for other purposes. (5) They collect some data solely in order to report them.

While school personnel have many different attitudes about this kind of work, it seems fair to suggest that if asked to rank data uses in order of importance to the legitimate functioning of a school, most administrators would rank them in the order presented in the previous paragraph. Maintaining student records is clearly a critical function, but as one descends the list one is more and more in the realm of "paperwork".

This ranking is critical in the school's perception of a computerized student data system. A data system which improves student record-keeping or day-to-day management is likely to be seen as a service, even if it does not actually save work. If the data system does nothing but improve report generation, it is likely to be seen as useless by nearly everyone. The only exception to this is if the system eliminates the need to complete a report which is presently produced and found to be obnoxious.

How can a data system either save work or improve the quality of a data activity? A data system can allow student data to be effortlessly re-used in many ways by performing such humdrum tasks as sorting, printing, and reformating. A data system can perform

more sophisticated information processing, such as test scoring, scheduling, and management of an applications/admissions system. A data system can compile statistical reports from individual data records. A data system can be an information resource, providing schools with needed information in the most convenient manner. Finally, a data system can improve data quality through editing, and by prompting schools for missing information. (This last "service" may not be perceived as such in a school with little regard for data quality.)

On the other hand, a data system can burden a school in a number of ways. School procedures may have to be performed by an inconvenient method, which has been mandated as a "simple" way of generating data for the data system. Schools may have to copy data onto input documents without deriving noticeable benefit from the computerization of these data. The data system may surreptitiously or openly lead to an increase in the total amount of data which schools must collect. And a data system may provide misinformation, or information so lacking in trustworthiness that the school should spend considerable time in verifying it before using it.

A student record system which is selected by a school administrator for school use is unlikely to be more burdensome than beneficial, and if it is, it will disappear quickly. If the school is not technically capable of handling the system, this will be apparent. Schools set their own standards for accuracy, timeliness, and completeness of data. If there are features of the system which are not well suited to school procedures or priorities, they can be modified. Whether the service is provided by a vendor or produced in-house on a small computer, the school can select a wide variety of options about data to be maintained, procedures to be followed, and outputs to be produced.

A data system like the ISDS cannot be tailored so closely to each school. The ISDS must support school record-keeping, but has two other functions as well: to promote the uniform maintenance of records on students who are mobile in the school system, and to improve the amount and quality of data available to administrators outside the school building. The individual school cannot be the judge of the adequacy of its procedures, nor can the individual school be given discretion over which data elements are maintained. No school can opt out of the system by refusing to submit data and corrections.

In its first few years, schools will inevitably see the ISDS as a set of external requirements. There is too little choice and not much data, and the data will seem of questionable accuracy to a school unless that school either submitted it or verified it. While everyone wants the schools to feel "ownership" in their share of the data system, it is clear that this will not happen right away.

One way of dealing with this problem has been to stress the usefulness to schools of the products which the system can produce. A complementary strategy has been to minimize the level of effort required of schools, by streamlining data collection activities so that they fit as closely as possible into school operations. These strategies have succeeded to the degree that services were genuine services (and not "thanks, but no thanks" outputs), and to the degree that the data collection instruments have been useful to the schools as documents (and not as the first stage in a long process with a report promised at the end). Appendix A presents a June, 1981 planning document which describes a number of possible products for 1981-82, many of which will have been produced by June, 1982. The rest of this section focuses on how to make data collection as unobtrusive and accurate as possible.

Designing accuracy into a data system

Forms design -- The more often a school consults and makes use of its ISDS data, the stronger the likelihood that the school will try to keep the data highly accurate. But the best assurance of accuracy would be for data system input documents to become the actual records kept by schools for their own use. Once schools come to rely on data system forms for their own recordkeeping, data submission becomes a by-product of internal processes, and not an external requirement.

From this point of view, it is the input documents rather than the system reports which are crucial. Schools may or may not wish to consult data system products, but they must have accurate records for their own use. If data system input forms are designed to be the actual school working records -- of which the data system periodically receives a copy -- then the difficulty of submission is minimized and the likelihood is great that the data will be as accurate as the school can make it.

The general models for such forms are familiar to data system designers. One is the individual record sheet, kept in a file. Another model is the preprinted list, with space for additional information about each person on the list. A third is a ledger for day-to-day recording of events or transactions. Whatever the form used, a copy is dispatched to the data center; if appropriate, a new set of forms is generated which display the updated information. Individual forms can be submitted whenever changes are made. Mass forms are sent at intervals or at specified times for collection of particular data elements.

For such a system to work, the forms must be in the schools prior to the recording of any data for the period of interest. The forms must be collected often enough so that multiple changes do not destroy legibility. The forms must be intelligible to school personnel and to a keypuncher, and they must be designed with in-school storage and access in mind.

Collection of new data -- This use of forms helps to combat three of the sources of error cited in Section II: failure to submit information, incorrect submission of information and failure to discover errors once made. It does not address the school's failure to collect the data in the first place. No method of data submission can solve a problem of data collection. Yet, central administrators frequently behave as though a data submission document will generate its own data collection procedure. In fact, for fear of causing offense to school administrators, central administrators rarely mandate or even suggest specific procedures for acquiring the data that must be submitted.

The reluctance to offend only makes sense where schools currently keep data well, so that all they are being asked to do is to implement new clerical procedures. If schools now keep the information badly, or have never kept the information before, then in-school data collection procedures must be developed and tested. This is quite clearly a central administrative role, although school personnel will be essential in the design and testing.

Ideally, these new internal procedures would be operational for some time before data are actually submitted to the ISDS. Schools would then have the opportunity to solve some of their data collection problems before the start-up of the new computerized data series. In addition, the mandate to assemble accurate data would be clearly distinguished from the mandate to submit the data to the computer.

Procedures for discovery and correction of error -- Whenever new data elements are to be added to the data system, a strategy for discovery and correction of error must be part of the initial design. This strategy can be refined on the basis of the results of pilot tests of forms. Any strategy must include computer edits for "impossible" data, a simple technology for the submission and entry of corrections, and the ability to monitor the entire process of error correction via summary reports.

In addition, the data must be examined for validity by an administrative office. No computerized system will automatically discover patterns of misunderstanding and misreporting. Unless a central office is interested enough in the data series to study it in detail, and to insist that errors be corrected, it is unlikely that all or even most schools will submit highly accurate data. If a data series is of such marginal interest that no central office will do this, then it is probably not appropriate to place the data element on the data system, except as a convenience to schools who wish to maintain the data.

Data which needs only to be "approximately correct" should be identified in advance and collected accordingly. Upgrading of quality can take place later. If all data elements need only be approximately correct, it may be inappropriate to add the data set to the ISDS; it can always reside on a separate file to be used as necessary.

Promoting Accuracy During Start-up and Operation

The collection of a data series for the first time usually requires that schools adopt new and somewhat unfamiliar procedures. It is true that schools have plenty of experience in responding

to surveys and in filling out forms, but standards of legibility, accuracy and completeness which suffice for most surveys may not be adequate for individual data records. One should not underestimate the amount of notice, training, and field supervision required, no matter how routine the task may seem.

In cases where the information-gathering is not routine -- if new data are to be collected, or if each student's record must be examined closely -- it is important to be sure that school staff can devote major attention to the work. The most hectic times during the year -- the entire month of September, much of October, and the end of the term -- are times when it is particularly difficult to collect accurate data unless schools have received sufficient advance notice to make room for the data collection in a very tight calendar.

In general, start-up is a perilous phase in the life of a data series. Data collections are likely to have unforeseen problems and complexities no matter how careful the plan and training. Adequate time must be allowed for retrieving data from recalcitrant schools, for returning unsatisfactory submissions, and for handling complications in data entry and processing. A correction round, in which schools must respond to errors discovered in processing, is also part of the original data collection.

Not until these errors are corrected and re-entered is the system ready for operation, or for generation of output reports. It is important not to promise reports too early. Should statistical information be required as a by-product of the data collection round, schools should submit summaries along with the individual data. These summaries are required anyway as control totals. Control totals are vital to ensure that all data are received and processed. They should be so defined as to be useful

to schools, by focusing attention on important attributes of the data. They should not be eliminated in a misguided effort to "save work" for the schools in order to justify the data system.

If it is possible to pilot test the collection, entry, and processing of the data with a subset of the schools, then the start-up may be simpler in the rest of the system. It is a good idea to look for "natural" pilots: cases for which there is a good reason to do the planning, forms design, programming and data collection with a reduced number of schools.

The addition of a single new data element to a pre-existing form and procedure should be regarded as a "start-up" activity in this sense of the section. While certainly such a start-up is simpler than the initiation of a new set of procedures, it still requires much attention if we are to avoid putting inaccurate data on the data system.

Once the start-up has established a good data series, it becomes essential not to let the quality decay with the passing of time. Schools must have the ability to update information on a continuous basis, and to make any corrections which appear warranted. Ideally, these updates will be entered as they happen or are discovered on a data entry terminal. But it is almost as satisfactory for schools to record updates on documents which can be submitted on a cycle which reflects the timing and volume of changes. As long as the schools' working records can be updated, school staff will be less frustrated by having to wait to submit changes to the data system.

This continuous in-school use of data system documents means that it is inadvisable to change basic documents except when absolutely necessary, and even then only with considerable notice

and preparation. A common error in maintaining a data system is to change procedures frequently and erratically. Certainly a data system must be flexible if it is to meet the need of the school system. Yet redesign and replacement of forms, communication of new procedures and changing of computer programs, all consume major amounts of energy, energy that in most cases would be better expended on new applications.

This does not mean that the data system should not be used to generate forms for ad hoc data collections or special reports. Nor does it mean that procedures that are not working should be kept. But it does mean that changes should be infrequent and introduced very carefully.

One final point. Despite the desire to let the data system enter the school system as unobtrusively as possible, it is important to face up to the fact that it will eventually cause a change in job definitions in the schools. Schools will need to have coordinators who take prime responsibility for data system submissions and activities whether or not they actually perform the work. This role is often performed by an assistant principal or a secretary, if it is performed at all; it may become desirable to standardize this assignment and to provide special training.

V. ENSURING COMPLIANCE

As discussed in the introduction to the previous section, schools perceive a mandatory data system very differently from a locally-developed or procured system. A system will be selected voluntarily because the principal anticipates that the assistance it will provide will be worth the cost and effort. The principal rarely makes this decision without one or several enthusiastic staff, who may actually have proposed that the system be procured or developed. In addition, the principal will retain control over record-keeping and reporting; the data system works for the principal and extends the principal's power.

A mandatory system must be implemented in every school whether or not the principal considers it to be worth the trouble, and whether or not the school has anyone on staff who is interested in controlling or using the data. In addition, while the principal may benefit from the information produced by the system, the school loses a great deal of control over its own information and its record-keeping practices. If the administrators of the data system attempt to maintain data quality--as they certainly should--then existing school procedures may be exposed as inadequate. For all these reasons, we may expect that some schools will be less than eager to comply with system mandates.

In a decentralized system like the New York City schools, the local superintendent has considerable operating authority and may be better able than central administrators to monitor performance on-site and to enforce change. The ISDS promises to strengthen the authority of Community Superintendents and High School Borough Superintendents by increasing the information they have available. However, the data system also weakens superintendents' control over the release of data about their districts, and some superintendents

may see the ISDS as a challenge to their authority rather than a useful tool.

We have stressed negative factors not because of widespread opposition and lack of cooperation. In fact, the vast majority of schools and districts have been remarkably responsive to the demands made on them, and forthcoming with suggestions as well. But even a few isolated cases of non-cooperation jeopardize the general trustworthiness of system data in a school system in which students are highly mobile as in New York. In addition, as this entire paper has suggested, the task of creating highly accurate data on all students is one which will require more than cheerful compliance with central directives.

In general, we find it helpful to distinguish among five levels of "participation" or "compliance" by individual schools:

(a) Acceptance of major responsibility for maintaining fully accurate computer files.

(b) Willingness to undertake reasonable effort and care, including supervision of all clerical activity by competent administrators (or secretaries, if appropriate).

(c) Willingness to complete all forms, but not to track down hard-to-find information or to resolve anomalies.

(d) Perfunctory and careless clerical work, at some or all times.

(e) Non-submission of data, or submission of flagrantly incomplete or inaccurate data.

This range of attitudes is not unlike the range a teacher finds in a classroom, and, like teachers, one must focus major attention on the low achievers, strive to raise the mean, and encourage the best performers while not designing the system around them.

The schools in category (e) are the most obvious threats to a useful system, but category (d) is almost as threatening. Schools which go through the motions of compliance are less likely to be identified as problem schools than are schools which refuse to cooperate; data from Category (d) schools, while inadequate, may still be accepted for processing. Thus a major priority for data system managers is to identify both categories of schools as quickly as possible.

Category (c) reflects a "business-as-usual" attitude. This attitude is likely to prevail in many schools for some time. These schools will submit data which are generally adequate for Level 3 uses, and through feedback, monitoring, and auditing, the performance (if not the attitude) of these schools can be raised to Level 2 standards. Until very few schools are in this category, the data system should not be used for functions which we characterized as Level 1B or 1A.

Category (b) is a worthwhile goal. The managers of the data system should attempt to create an attitude in the schools in which data system requirements are accorded the respect and status due to a major day-to-day administative function.

Until many or most schools are in Category (b) there is little point in expecting the increased responsibility of Category (a). Schools which labor to produce "perfect" data will have to deal with inadequate information about incoming students, and it is hard to argue that the data system products will be worth their level of

effort until nearly all schools are providing very good data. However, as we discussed earlier, some stellar performers are bound to emerge, and the system should make use of them in a variety of ways, particularly in the design of new components, forms, and training materials.

Specific recommendations for ensuring compliance

The most important factor in compliance is the creation of an intelligent and useful system that works. Most school personnel will comply with almost anything, no matter how absurd, provided that it comes through proper channels. As professionals, they will give extra effort to an activity that seems to have real purpose and to be likely to accomplish its goals. Accurate data and minimal need for correction are strong incentives for doing things right the first time.

The second most important factor in compliance is the monitoring of performance. If people think that you are keeping track of how well they are doing, they are far more likely to do well. A straightforward system for evaluating, recording, and reporting the timeliness and adequacy of each submission of each school is the key to performance monitoring. Such a system should be based on logs, on manual inspection, and on data system reports.

Non-compliant schools should be identified immediately and dealt with through formal channels. This will be acceptable provided that the administrators of the data system make it their business to be aware of extenuating circumstances, and to be willing to devote data system resources to cases that need help. Data system administrators should also be aware of whether a school has been compliant in the past, as it is counter-productive to

treat each collection as a unique activity.

A major virtue of a manual (or automated) evaluation system is that it assists you in distinguishing non-compliant schools from partially compliant schools, and in recognizing extenuating circumstances when they happen. In a small system, this is not a problem, as each school quickly becomes a unique entity to the data system staff. In a system with close to 1000 schools, this personal knowledge of individual schools can be expected only of field staff, and even of them only with time.

In addition to these monitoring activities, a program of auditing is essential in assuring that schools report data which is consistent with other school records and with actual circumstances. The audit activity should include in-house comparisons of data submissions to other data which purport to measure something similar (e.g. survey data, prior year submissions), as well as a program of field auditing. The field auditing may or not be connected to a program of field supervision and training.

In designing a field audit, there are two objectives which conflict. On the one hand, the audit seeks to establish the general level of accuracy of the system so that system data can be presented along with estimates of error. On the other hand, auditing is a compliance mechanism which should focus on schools which are suspected to have sent in faulty data. A sophisticated reseach design can meet both those objectives, but has the flaw of requiring weighting to come up with system-wide estimates of accuracy. Perhaps the best alternative is to have a "system audit" for generatng estimates of system accuracy, supplemented by a "compliance audit", whose results would not be used in generating aggregate estimates. This is a highly inefficent research design, but auditing is not only a "research" activity aimed at measurement

of average error. In fact, a major purpose of the compliance audit would be to gain greater insight into the reasons for non-compliance and possible need for re-design.

Once an audit mechanism is in place, it becomes possible to use the data system for such Level 2 functions as funds allocation and monitoring of compliance with administrative and educational policies. These uses of the data are themselves very strong incentives for accuracy, but it is important not to institute them until the general error rate due to misunderstanding of data system procedures is low. Otherwise the policy monitors will find themselves to be spending most of their time on resolving clerical errors, and the data system may be perceived as inaccurate and misleading.

In general, schools should have both intrinsic rewards for compliance (better and therefore more useful data) and extrinsic rewards (a good record on management reports and audits). These are the heart of any strategy for assuring compliance.

Finally, as a last incentive for compliance, there might be some specific rewards for highly cooperative schools, such as first-in/first-out processing, and priority for special requests and training.

VI. THE ROLE OF THE DATA SYSTEM IN PROMOTING DATA ACCURACY--A FINAL NOTE

The entire paper up to this point has conceptualized the data system and its managers as supervisors, whose task is to "make" schools perform up to an acceptable standard. This way of posing the problem does not mean that we believe that schools should regard themselves as subservient to the data system managers.

Nothing about this paper suggests that data system policy or standards must be made centrally, rather than by a representative committee of users. This paper simply suggests what must be done if the data system is to meet its assigned responsibilities with adequate protection for individuals and their records.

In addition, this paper has very extensive implications for the internal management of a data system. While the schools are responsible for submitting accurate data, the data system designers are responsible for creating a framework in which this task is perceived as reasonable rather than onerous, useful rather than pointless, and serviceable to the field rather than a further extension of central control. The data system designers must construct an internal system of record-keeping and control, as well as a powerful field liaison staff which can assist in the job of tailoring a data system to the complexity of the school system.

Above all, in shifting some responsibility away from principals in assuring data quality, that responsibility is shifted elsewhere. Data system staff will be asked questions never before asked, and they will be told to give answers which are not hedged by uncertainty about the accuracy of data submitted by schools. The New York City school system, through its Chancellor, has decided that such a resource be made available. The agency responsible for carrying out this mandate

must somehow see to it that the data do meet the standards required of them. In a decentralized, enormous school system, this will be quite a challenge.

APPENDIX A

PROPOSED PRODUCTS, ACTIVITIES AND RELATED BENEFITS OF THE STUDENT BIOGRAPHICAL SYSTEM, 1981-82 SCHOOL YEAR

I. School Administration

Benefit	Product or Activity
Reduce or eliminate the need for the repetitive manual transcription of data for use in internal school organization, e.g. classroom and teaching assignments. Without duplication of effort flag individual students for referral to guidance counselors and attendance officers based on standardized test score results, attendance or mobility data. Assist school administrators in future planning by providing longitudinal data similar to the above.	Reference rosters, both summary and detail: -alpha -grade -class -longitudinal studies, i.e admissions and discharges
Eliminate the need to hand-address envelopes for mailing parental notifications and other materials to pupils' homes.	Mailing labels
Eliminate the need to transcribe updated biographical data, e.g., change in address, admission or discharge data, onto multiple sets of school record cards.	Labels for placement on record cards
Eliminate the need for handwritten office cards, class cards, and other internal records for incoming students.	Computer-generated cards
Introduce the beginnings of a standardized pupil accounting system in schools which is responsive to schools' needs, yet is readily understood by newly assigned school secretaries and other school personnel. This is especially important due to the high turnover of secretaries and use of part-time personnel.	Replacement of 209A and B with admission and discharge forms in lieu of U-forms. Monthly reconciliation reports of admissions and discharges and summaries of registers by grade and class
Substantially reduce or eliminate the clerical task of completing biographical data on competency test answer sheets and certain other tests, while improving the accuracy of score reporting.	Pre-slugged answer sheets for January and June RCT's, Spring PSEN's and Citywide examinations

SOURCE: New York City Public Schools, Office of Student Information Services, July, 1981

Benefit	Product or Activity
Articulate RCT test scores for incoming high school students who were tested in junior high or intermediate school in the Spring.	Score rosters and labels will be sent to the receiving H.S.'s this Fall
Early notification to receiving schools of long-term absence and immunization status of incoming students.	Expanded list notice procedure provided this data to receiving schools in the Spring. Rosters with similar data will be be made available to Superintendants in the Fall
Eliminate the need for UAPC H.S.'s to manually report RCT Math and Reading and PSEN test scores to UAPC.	O.S.I.S. provides data tapes directly to U.A.P.C.

II. Community School Districts/High School Superintendents

Benefit	Product or Activity
Provide hard-copy student biographical, testing and other data for planning, resource allocation, budgeting, reconciliation, audit and other administrative and managerial tasks without burdensome clerical work on the part of schools. Several of such requests were received and fulfilled this Spring.	Reference rosters, both summary and detail: -alpha -grade -class -long term absence -longitudinal studies, i.e. admissions and discharges
Provide computer-accessible data for input to Districts' computerized systems and to vendors hired by districts to provide computerized services without the need for duplicative work on the part of schools.	Computer tapes of subsets of the Biofile, custom tailored for users' needs

III. Central Offices

Benefit	Product or Activity
Provide hard-copy student biographical, testing and other data for planning, resource allocation, budgeting, reconciliation, audit and other administrative and managerial tasks without burdensome work on the part of schools. Examples include rosters of "lost children" (i.e. students discharged to a new school who fail to appear at that or any other school) and dropout/discharge reports.	Reference rosters, both summary and detail: -alpha -grade -class -long term absence -longitudinal studies, i.e. admissions and discharges

Benefit	Product or Activity
Provide computer-accessible data for central automated systems on an on-going basis, without need for duplicative work on the part of schools. Of particular importance is the need to reconcile monthly admission and discharge data amongst central administrative systems, such as those for the Bureau of Pupil Transportation and the Division of Special Education	Computer tapes of subsets of the Biofile, custom tailored for users needs
Provide hard-copy and computer-accessible data for survey purposes, including the pre-printing of data collection forms, and elimination of additional data collections where possible. Examples include the provision of the Impact Aid address file, School Profiles register counts, data for Occupational Education surveys and a field audit instrument for the October 31 Audited Student Registers	Rosters and computer tapes, as indicated above

APPENDIX B

BIBLIOGRAPHICAL NOTE

The Plan for an Integrated Student Data System (1981) was prepared under a grant provided by the New York State Education Department under Title IV-C of the Elementary and Secondary Education Act. The grant funded the Planning and Analysis Unit of the Office of Student Information Services: Charles Troob, Susan Choy, Joseph Reswick, and Susan Ryzewic.

The unit was assisted by two consultants provided by I.B.M., Susan Leslie and James Vacirca. Edward Pyle of the Office of Data Processing also provided assistance. General superivision was provided by Thomas Gross, Director (Acting) of the Office of Student Information Services.